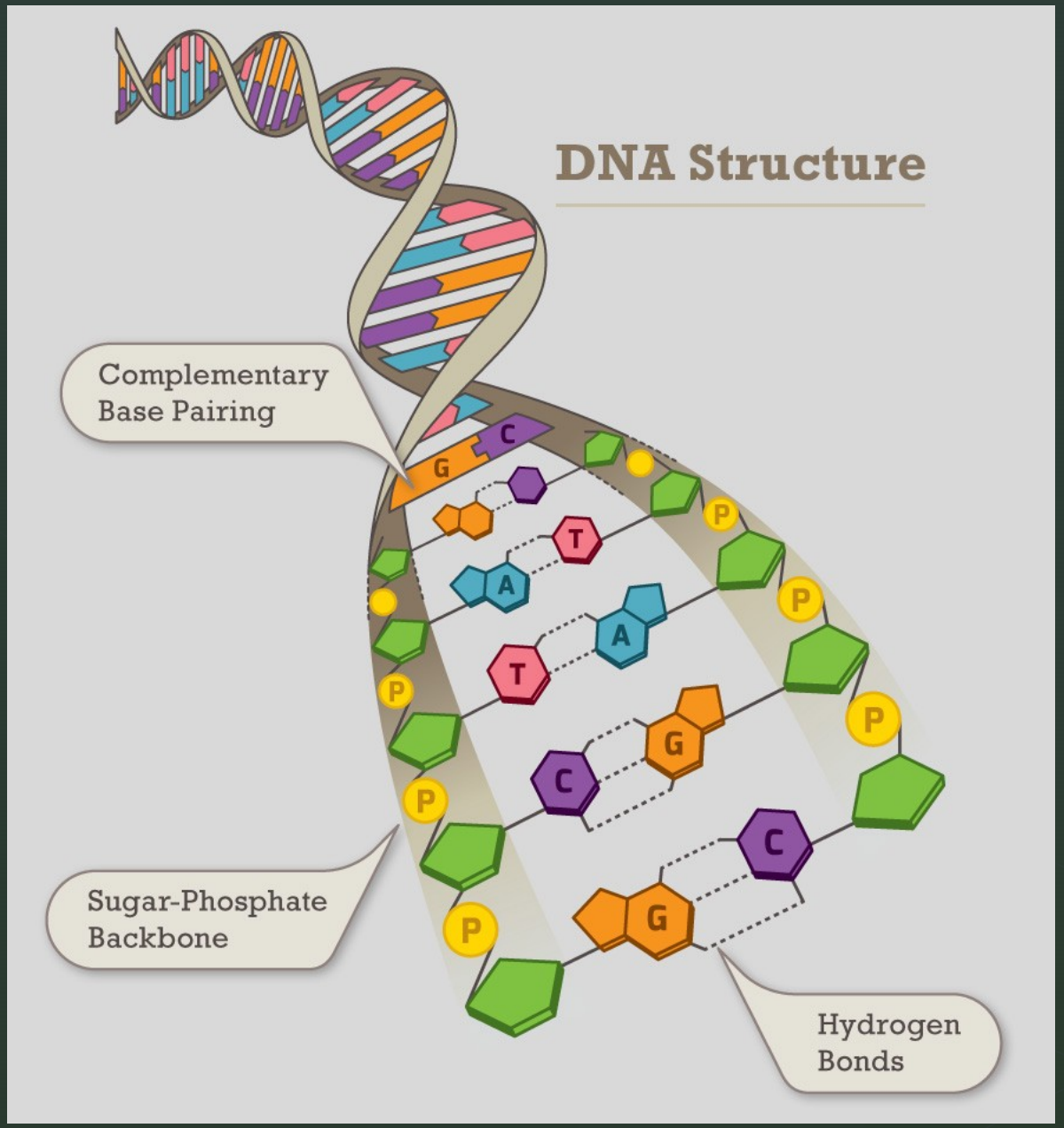
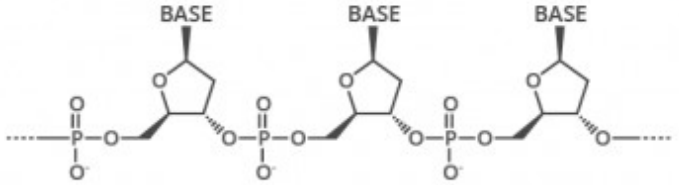
# NEXT GENERATION SEQUENCING DATA ANALYSIS


DNA Structure
Complementary Base Pairing
Sugar-Phosphate Backbone
Hydrogen Bonds
C
G
T
A
A
T
G
C
C
G
P

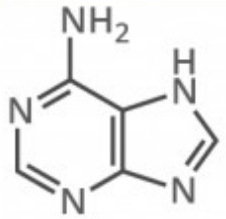
# THE CHEMICAL STRUCTURE OF DNA

## THE SUGAR PHOSPHATE 'BACKBONE'

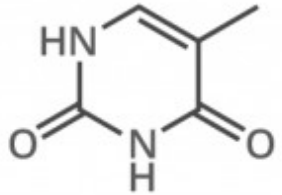



DNA is a polymer made up of units called nucleotides. The nucleotides are made of three different components: a sugar group, a phosphate group, and a base. There are four different bases: adenine, thymine, guanine and cytosine.

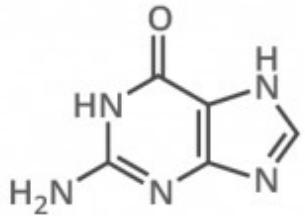
### A ADENINE



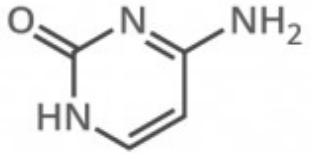
### T THYMINE



### G GUANINE

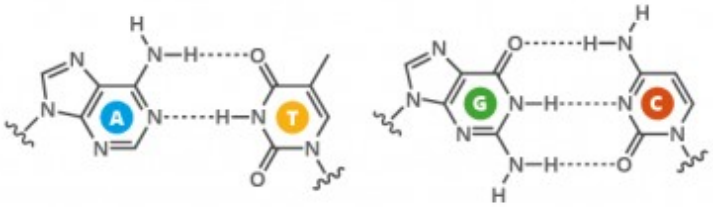


### C CYTOSINE



## WHAT HOLDS DNA STRANDS TOGETHER?

DNA strands are held together by hydrogen bonds between bases on adjacent strands. Adenine (A) always pairs with thymine (T), while guanine (G) always pairs with cytosine (C). Adenine pairs with uracil (U) in RNA.



## FROM DNA TO PROTEINS

The bases on a single strand of DNA act as a code. The letters form three letter codons, which code for amino acids - the building blocks of proteins.

DNA → TRANSCRIPTION → RNA → TRANSLATION → PROTEIN

An enzyme, RNA polymerase, transcribes DNA into mRNA (messenger ribonucleic acid). It splits apart the two strands that form the double helix, then reads a strand and copies the sequence of nucleotides. The only difference between the RNA and the original DNA is that in the place of thymine (T), another base with a similar structure is used: uracil (U).

DNA SEQUENCE T T C C T G A A C C C G T T A
mRNA SEQUENCE U U C C U G A A C C C G U U A
AMINO ACID
Phenylalanine
Leucine
Asparagine
Proline
Leucine

In multicellular organisms, the mRNA carries genetic code out of the cell nucleus, to the cytoplasm. Here, protein synthesis takes place. 'Translation' is the process of turning the mRNA's 'code' into proteins. Molecules called ribosomes carry out this process, building up proteins from the amino acids coded for.

© Andy Brunning/Compound Interest 2018 - www.compoundchem.com | Twitter: @compoundchem | FB: www.facebook.com/compoundchem
This graphic is shared under a Creative Commons Attribution-NonCommercial-NoDerivatives licence.

# Terms

- **Template:** A DNA/RNA sequence part of which is sequenced on a sequencing machine or assembled from raw sequences.
- **Segment:** A contiguous sequence or subsequence.
- **Read:** A raw sequence that comes off a sequencing machine. A read may consist of multiple segments. For sequencing data, reads are indexed by the order in which they are sequenced.

z

- **Linear alignment:**
- An alignment of a read to a single reference sequence that may include insertions, deletions, skips and clipping, but may not include direction changes (i.e., one portion of the alignment on forward strand and another portion of alignment on reverse strand). A linear alignment can be represented in a single SAM record.
- **Read alignment:** A linear alignment or a chimeric alignment that is the complete representation of the alignment of the read.

z

- **Multiple mapping:** The correct placement of a read may be ambiguous, e.g., due to repeats. In this case, there may be multiple read alignments for the same read. One of these alignments is considered primary.
- All the other alignments have the secondary alignment flag set in the SAM records that represent them.
- All the SAM records have the same QNAME and the same values for 0x40 and 0x80 flags.
- Typically the alignment designated primary is the best alignment, but the decision may be arbitrary.

# PURPOSE OF EFFICIENT TECHNIQUE?

- Conventional genetic diagnostic techniques increase the difficulty in mutation identification.
- New technology of next generation sequencing has put a great boost in this regard. It is an extremely powerful technique for such heterozygous disorders.
- The need of time is to develop a comprehensive **diagnostic test to screen** all treatable genetic disorders, followed by targeted NGS would provide an early detection of genetic disorders in a single go, this could save thousands of live and can provide timely treatment/assistance for many genetic disorders in future.

# Linkage Analysis

- The human genome is very large and complex having thousands of genes spanned through 23 sets of chromosomes. Therefore finding a disease gene is always been tricky.
- Linkage analysis is a traditional approach to find a disease gene. Linkage analysis is a powerful, family-based approach of developing a relationship between the loci i.e. two loci on the same chromosome are linked or not.
- In meiosis homologous chromosomes exchange segments as the basis for crossing over or recombination. The original arrangements of alleles on the two chromosomes are called the **parental combinations** whereas the new combinations that are formed due to crossing over are known as **recombinant**s.
- If two loci are physically close to each other on the same chromosome then there are rare chances that they will be separated by a recombination event.

# Homozygosity Mapping

- Homozygosity mapping is a powerful tool to map autosomal recessive gene defects. The method takes advantage of the fact that inbred affected individuals are likely to have two recessive copies of the disease allele from a common ancestor; i.e. two **identical-by-descent (IBD**) alleles.
- It has been proven to be useful in locating the genetic defects in both consanguineous and non-consanguineous (unrelated) unions.
- Since small chromosomal regions tend to be transmitted as a whole, affected individuals will also have identical-by-descent alleles at markers located nearby the disease locus and thus will be homozygous at these markers.

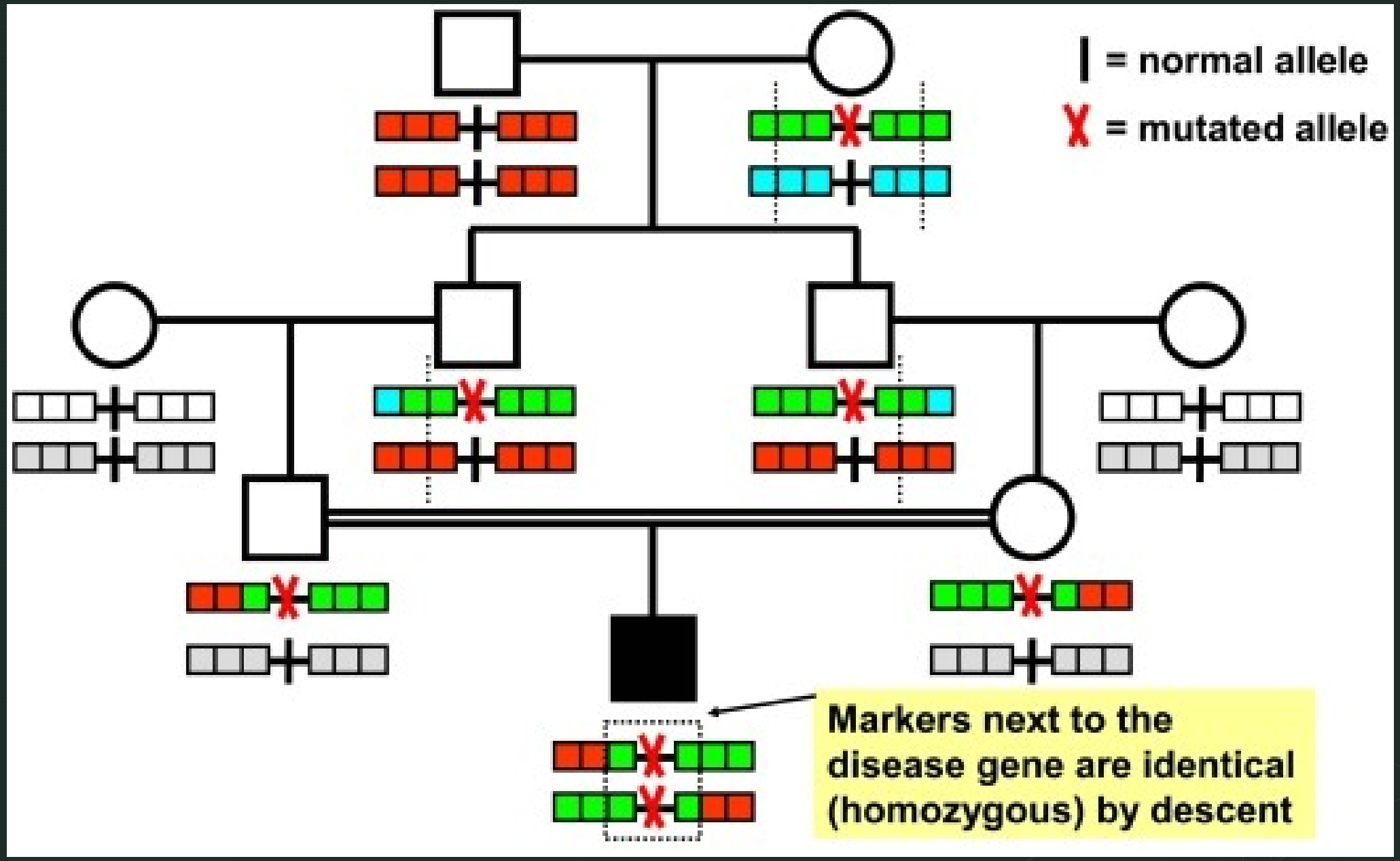
z

Homozygosity mapping takes advantage of the fact that not only the mutation but also the surrounding region is identical on both chromosomes.

These regions usually encompassing multiple variable DNA markers such as **microsatellite (STR)** markers or **single nucleotide polymorphisms (SNPs).**

By using these DNA markers **regions of homozygosity (RoH)** can be determined.

Therefore, this technique is called as **homozygosity mapping, IBD mapping or autozygosity mapping.**


= normal allele
= mutated allele
Markers next to the disease gene are identical (homozygous) by descent

# DNA Polymorphism—A Tool for Linkage Analysis

- The simple sequence repeats also known as **microsatellite**, have revolutionized the world of genotyping.
- STRs are hyper variable tandem sequence repeats which consist of di-tri- or tetra-nucleotide repeats.
- The most widely used STRs to be developed for genotyping are the simple (CA)n and (GT)n repeats. The (CA)n repeats are extremely abundant and can be found, on average, once every 30-60 kb. (CA)n repeats are generally polymorphic if the repeat length is greater than 10.
- By isolating and sequencing DNA fragments containing the microsatellite, PCR primers that flanked the STRs can be created and used to amplify it.

# Microarray-based SNP

- Microsatellite genotyping was replaced by a high throughput and robust microarray-based SNP typing, which can also detect CNVs.
- There are several platforms to perform SNP microarrays, offered by several companies such as Affymetrix (Santa Clara, CA) and Illumina (San Diego, CA) and others.
- They are high throughput, time and cost effective, useful to map both consanguineous and non-consanguineous families, and highly suitable to detect small heterozygous and homozygous CNVs.
- In addition, these dense SNP arrays are also helpful for genome-wide association studies (GWAS) to find the disease predisposition alleles in mutifactorial disorders like diabetes and cancer.

# Next Generation Sequencing

Sanger sequencing has been the standard technique from decades. This classical method is still the gold standard procedure in genomic diagnostic laboratories.

But it has some limitations like it has a limited potential,

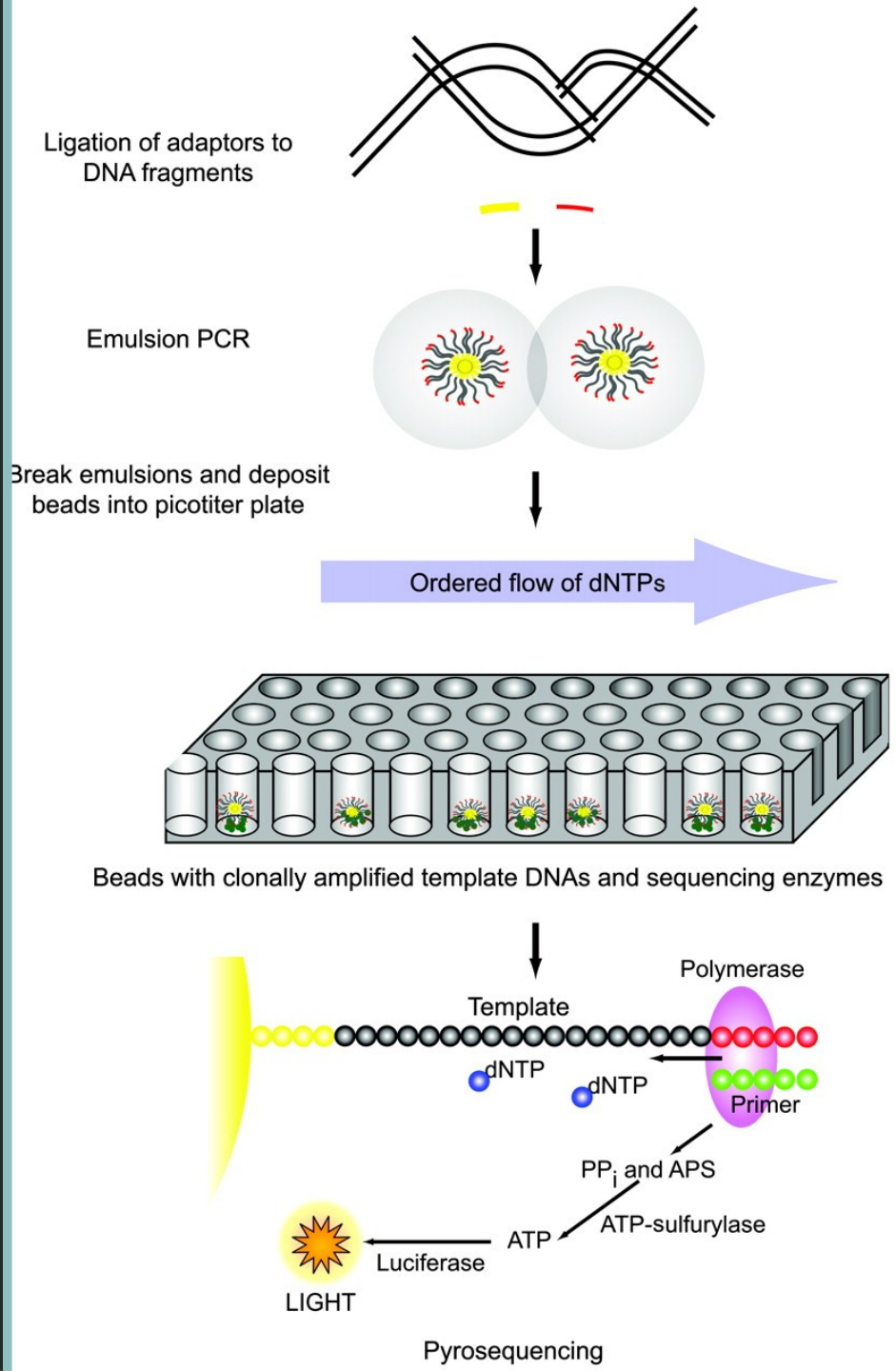
- maximum of 1000 nucleotides of DNA can be sequenced in a single run
- relatively low capacity (6 to 384 reactions in parallel can be handled)
- labor-intensive and
- expensive (costs $500/Mb of DNA)

# Next Generation Sequencing

- In the last few years, conventional Sanger sequencing is rapidly being surpassed by massively parallel sequencing (MPS), also called next generation sequencing (NGS) or second generation sequencing approaches.
- In past 4 years a large number new genes/ loci have been reported for autosomal recessive genetic diseases, and this was just because of the robust NGS technologies, which can massively generate hundreds of megabases to gigabases of DNA sequence in a single run depending on the type of platform.

z

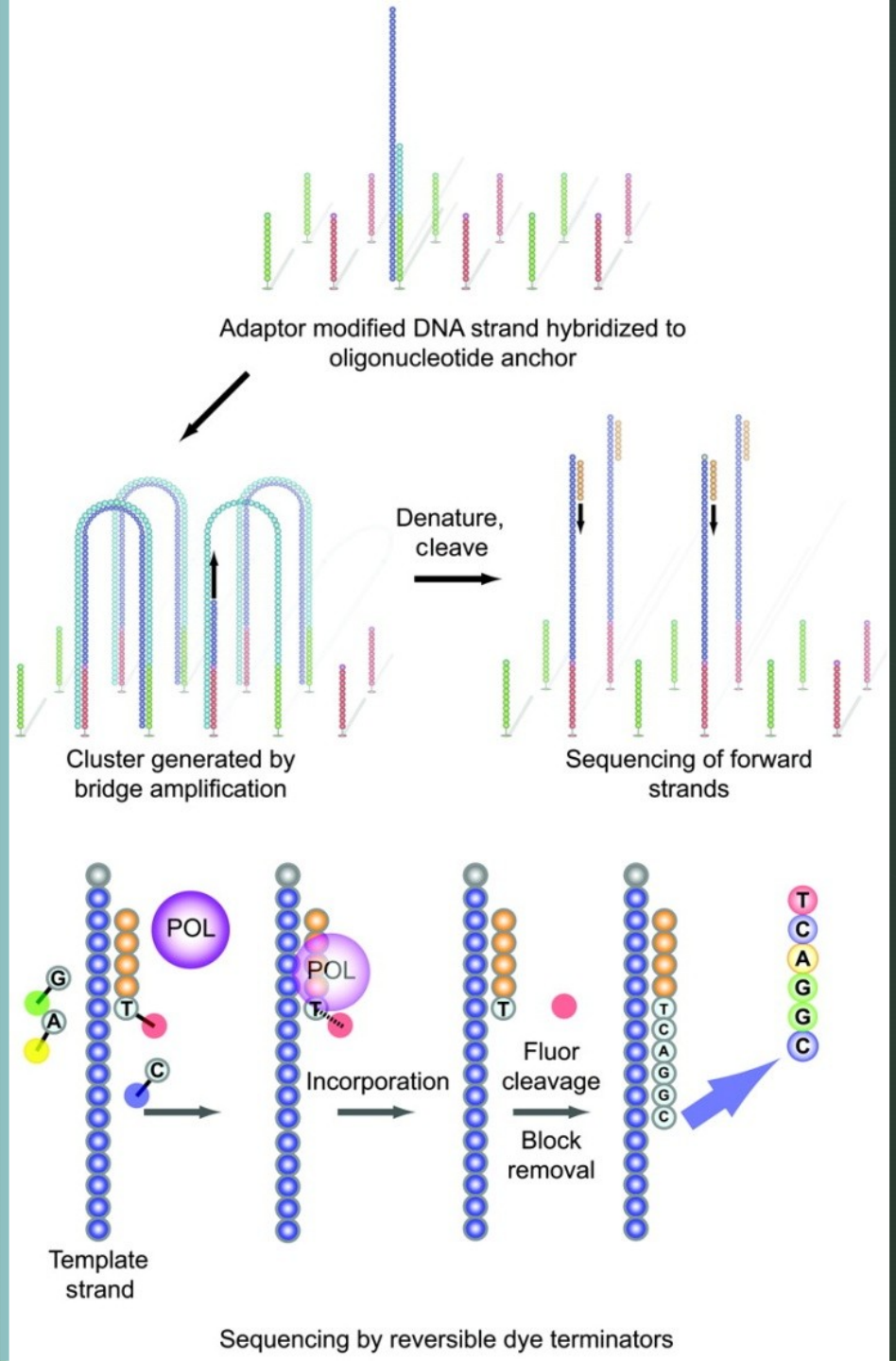
- Let’s have a look at different platforms available for next generation sequencing analysis.




# Roche 454 GS FLX sequencing

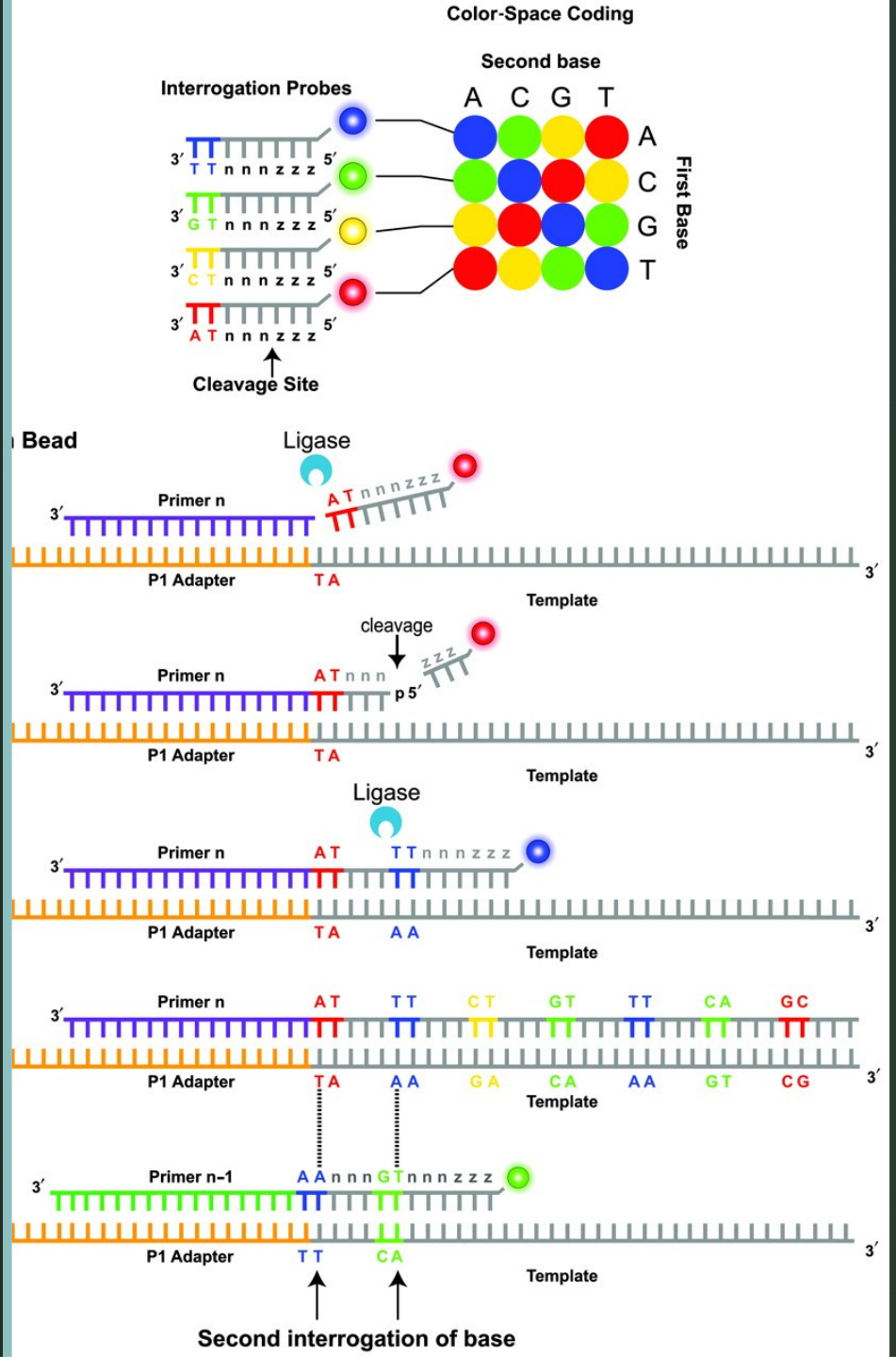
- Template DNA is fragmented, end-repaired, ligated to adapters, and clonally amplified by emulsion PCR.
- After amplification, the beads are deposited into picotiter-plate wells with sequencing enzymes.
- The picotiter plate functions as a flow cell where iterative pyro-sequencing is performed.
- A nucleotide-incorporation event results in pyrophosphate ($PP_i$) release and well-localized luminescence.
- APS, adenosine 5′-phosphosulfate.

# Illumina Genome Analyzer and HiSeq



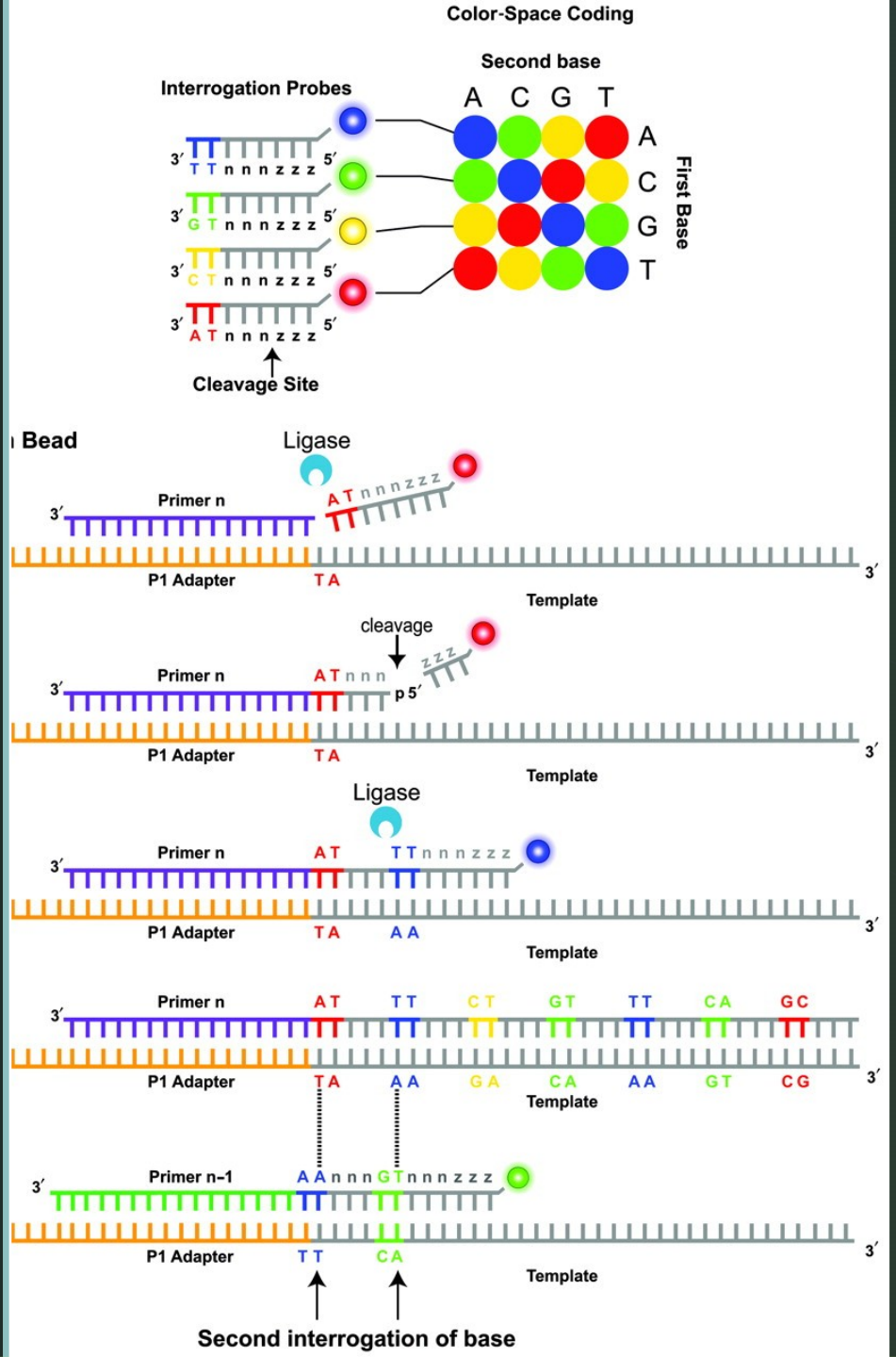

- Adapter-modified, single-stranded DNA is added to the flow cell and immobilized by hybridization.
- Bridge amplification generates clonally amplified clusters.
- Clusters are denatured and cleaved.
- Sequencing is initiated with addition of primer, polymerase (POL) and 4 reversible dye terminators.
- Post-incorporation fluorescence is recorded. The flour and block are removed before the next synthesis cycle.

# Applied Biosystems SOLiD sequencing by ligation




- SOLiD color-space coding. Each interrogation probe is an octamer, which consists of (3′-to-5′ direction) 2 probe-specific bases followed by 6 degenerate bases (nnnzzz) with one of 4 fluorescent labels linked to the 5′ end.
- The 2 probe-specific bases consist of one of 16 possible 2-base combinations.
- (A) The P1 adapter and template with annealed primer (n) is interrogated by probes representing the 16 possible 2-base combinations. In this example, the 2 specific bases complementary to the template are AT.

# Applied Biosystems SOLiD sequencing by ligation




- (B) After annealing and ligation of the probe, fluorescence is recorded before cleavage of the last 3 degenerate probe bases. The 5′ end of the cleaved probe is phosphorylated (not shown) before the second sequencing step.
- (C) Annealing and ligation of the next probe.
- (D) Complete extension of primer (n) through the first round consisting of 7 cycles of ligation.
- (E) The product extended from primer (n) is denatured from the adapter/template, and the second round of sequencing is performed with primer (n − 1).
- With the use of progressively offset primers, in this example (n − 1), adapter bases are sequenced, and this known sequence is used in conjunction with the color-space coding for determining the template sequence by deconvolution.
- In this technology, template bases are interrogated twice.

z

- All of these platforms have more or less similar workflows but they differ in other aspects such as sequencing chemistries

| | Roche 454 GS FLX | Illumina Genome Analyzer | Applied Biosystems SOLiD | Sanger |
|---|---|---|---|---|
| Sequencing method | Pyrosequencing | Reversible dye terminators | Sequencing by ligation | Dye terminators |
| Read lengths | 400 bases | 36 bases | 35 bases | 800 bp |
| Sequencing run time | 10 h | 2.5 days | 6 days | 3 h |
| Total bases per run | 500 Mb | 1.5 Gb | 4 Gb | 800 bp |

# Pros of NGS

- NGS can be used to sequence the entire genome of an individual, referred to **as whole genome sequencing (WGS),** as well as only exons of the genome which is referred to **as whole exome sequencing (WES) or exome sequencing**.
- During the last ten years, exome sequencing has been the most frequently used tool in research and diagnostics, as most of the disease causing mutations reside in the coding part and it is cheaper.
- NGS technologies has the advantage over other techniques as it helps in identifying disease causing genes regardless of DNA variations type, phenotypic, genetic heterogeneity and inheritance patterns.
- NGS is used to identify point mutations, small insertions or deletions (indels) and CNVs.

# Cons of NGS

- Despite the successes of exome sequencing, there are also some shortcomings in this methodology.
- In gene identification process exome sequencing cannot be used as the only tool, as all of the exons are not covered.
- First exon is mostly missing because of the high G-C content.
- In addition, repetitive sequences are a big problem in the process of fragment alignments.
- Furthermore, the depth of the coverage of all the exons is not uniform across the exomes and non-coding genes, such as micro-RNAs and other non-coding RNAs, are hardly covered by present enrichment kits.

# Future of Sequencing

- The development of novel sequencing technologies proceeds at a high pace.
- These novel technologies would overcomes the problems of current technologies and will have
  - higher throughput,
  - longer read length,
  - higher consensus accuracy,
  - small amount of starting material,
  - low costs and
  - faster throughput.
- Having all these advantages in sequencing technology, it is anticipated that there will be another rapid boost-up in near future in the field of genetics and molecular biology research.

# Conclusion

- Before next generation sequencing came into research, only a few genes were known to be involved in genetic diseases.
- By the use of NGS and specially exome sequencing gene identification has become so rapid that with every passing month a few new genes came on the scene.
- There are many large scale projects going on in different laboratories of the world, that have major concern to solve unexplained forms of genetic and developmental disorders by the use of next generation sequencing, also the medical genetics are using exome sequencing as first big test to solve genetic problems.

z

- THANK YOU